AFOSR-70-2883TR

COOPERATION / CONFLICT RESEARCH GROUP

Approved for public release;
distribution unlimited.

70-5

A Division of Computer Institute
for Social Science Research
Michigan State University

60

A SEQUENTIAL BAYESIAN MODEL

EVALUATION TECHNIQUE APPLIED TO

A MODEL OF CONCEPT IDENTIFICATION

By

Gary H. Mendelsohn

CONTRACT #
F44620-69C-0114

Report 70-5

Cooperation/Conflict Research Group

Michigan State University

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Evaluating mathematical models is currently a major concern of psychology. The question of "How good is the model?" can be answered in many different ways. Most commonly, statistical tests are generated to compare various aspects of the data with predictions from the model. If many of these tests fail to reject the hypothesis that the data is the same as the predictions, the model is considered tenable. A typical example of this is the performance of chi-square tests on theoretical versus obtained data; the data are usually something like averaged learning curves for groups of subjects.

The traditional method of evaluation described above does give information about the fit of the model. This method has heuristic value in model construction, in that the aspects of the data analysed are usually closely related to specific axioms in the model being tested. This characteristic gives evidence for the acceptability of individual postulates; this information is useful in revising a model at the axiomatic level. However, much information and detail is lost from the data in the process. The statistical tests are often performed on reduced data; averages over subjects may cancel out many differences which really occurred; and finally, the traditional method seldom gives any indication of the fit of the model to individual subjects.

It is possible to develop an alternative or alternatives to the traditional procedures of evaluation mentioned above. Having a procedure produce a single goodness of fit index for the model would be a good requisite. Another possible requirement could be that the data not be reduced by averaging or ignoring parts of it. A third requirement (or restriction) could be that the evaluation apply to the fit of the model to the individual, which allows many more parameters to be assumed constant over multiple data points.

The alternative technique of model evaluation developed in this paper is best described as a sequential Bayesian evaluation of model fit to individuals. An overview of the method starts with a prior probability that a given subject behaves according to the model. This probability is modified by applying Bayes' Theorem with the subject's whole protocol. The resulting posterior probability is then treated as a prior probability and Bayes' Theorem is again applied using the subject's next protocol. The sequence is repeated through all of the data, resulting in a final conditional probability of the model given the data.

The use of Bayesian concepts of probability is of great value in a behavioral science. It cannot be denied that oftentimes, intuition is the best predictor of behavior. The prior probability of Bayes' Theorem allows the experimenter's intuition to validly enter the evaluation process. Edwards, Lindman, and Savage (1963) put forth a rationale for the use of a Bayesian philosophy of statistics by the psychological researcher. They mention the ability of a Bayesian hypothesis test to actually accept an hypothesis directly, as opposed to the traditional acceptance by rejection of the null hypothesis. They also discuss the "limbo" state of an hypothesis when the null hypothesis is not rejected by traditional statistics. In closing, Edwards, et. al., make the general comment that "the Bayesian outlook is flexible, encouraging imagination and criticism in its everyday applications" (1963, p.240). This attitude is good in that it overcomes the tendency -- with traditional statistics -- to use numerous but irrelevant statistics to lend respectability to otherwise tenuous conclusions.

It might be argued that an irresponsible experimenter could begin his evaluation with an inordinately large prior probability and collect a small amount of data which did not greatly affect this probability. This is true, but the responsible experimenter who chooses a reasonable prior probability and collects a fair amount of data will find that his final probability of the model given the data will closely resemble any other responsible experimenter's conclusions based on the same data. This convergence of Bayesian "opinions" after sufficient data collection, has been proven (Blackwell and Dubins, 1962).

To implement the present sequential Bayesian evaluation technique, three things are necessary. All parameters of the model must be estimated for each subject. Secondly, a probability expression must be derived from the model for any possible protocol. The third requirement is the probability distribution of possible protocols which could result from the task. Given these three necessities, it remains only for the researcher to substitute his prior probability and the data in Bayes' formula to arrive at a single number which is the conditional probability that the subject behaved according to the model, given the data. In a sense, this probability is the probability that the model is true -- a kind of absolute evaluation that is not found in traditional model testing techniques.

The current sequential Bayesian approach to model evaluation has the desireable characteristic of using all of the information available in the data. In addition, its straightforward approach is applicable to any well defined mathematical model. The sequential Bayesian technique lacks the heuristic side benefits of yielding independent tests of individual axioms of the model, but the technique may be applied rapidly to many different models in order to determine the effects of changes in the initial model. The researcher, working with the types of models to which this evaluation would apply, would know the model well enough to modify it without the direct information about which postulates need changing that a traditional evaluation might make available.

The outline of the sequential Bayesian model evaluation technique presented above will be expanded and the technique will be applied in the rest of this paper. The application will be an experimental test of a concept identification model.

Concept identification is a fundamental behavioral process. It can be defined as a cognitive process for determining the classification rules for perceptual objects. As a basic process of behavior, concept identification needs investigation and specification, perhaps through mathematical modeling. Good models of concept identification behavior could form part of the foundation for a well defined, organized, and useful science of behavior.

PROBLEM

Many factors are involved in basic processes such as concept identification. The experiment described in this paper was intended to begin a list of factors which affect concept identification behavior. Two of the many possible factors were selected for investigation. The nature of the perceptual object was varied within boundaries established in the design of the experiment. The boundaries established for perceptual objects were that they be visual stimuli. Simple geometric designs were selected; they were designed to carry information on only four dimensions. Each dimension was to be binary -- it could take on only one of two values. The way in which the nature of the perceptual object was varied was to present either a single design, or a double design in which the second half of the design was completely redundant with the first half.

The second effect to be examined was the nature in which the subject's behavioral task was defined for him. The task definition was varied by instructing subjects in different ways with different intents.

Only actual concept identification behavior was considered; the acquisition of that behavior was excluded from consideration. A pilot study indicated that following instructions and three practice tasks, subjects appeared to be stable in their strategies of concept identification behavior.

The experiment reported here was designed then, to have subjects display concept identification behavior; instructions and method of stimulus presentation were varied.

The Bower and Trabasso (Atkinson, Bower, and Crothers, 1965) model of concept identification was selected to demonstrate the sequential Bayesian evaluation technique. The principles of this technique required that a suitable estimator of the model's parameter be derived and applied; that the expression for the probability of any protocol be derived; and that the distribution of protocols under any possible model be estimated.

In summary, the problems addressed in this paper are the definition of factors affecting concept identification behavior, and the derivation and application of a sequential Bayesian model evaluation technique to the Bower and Trabasso model of concept identification.

METHOD

Subjects

The subjects used in this experiment were volunteers from an introductory psychology course taught at Michigan State University. It was required that each subject be naive concerning experiments similar to the present one and that each subject participate only once. Due to the availability of qualified subjects, all subjects used were female. Approximately one hundred subjects were run during the course of the experiment. Some of these gave evidence of not understanding the instructions after having started the experiment. A few were unable to complete all the experimental tasks within the alloted time. For a small number of subjects, experimenter error or apparatus failure occured. The final number of subjects used in the data analyses was seventy-four. The composition of the four experimental groups is described in Table 1. The labels of the four groups will be defined later.

Table 1. Subject characteristics.

Group	1S	1L	2S	2L
Freshman	10	15	15	17
Sophomores	5	4	3	2
Juniors	1	0	1	1
Total	16	19	19	20

Apparatus

This experiment was fully automated with the exception of the instructions, which were read aloud by the experimenter while the subject read along from a typed copy.

The subject was seated at a table facing a brown Masonite panel measuring 24 inches high by 30 inches wide. In the lower center of this panel was a white, opal glass screen measuring 8 inches high by 12 inches wide; the lower edge of the screen was about 5 inches from the table top. The stimuli and reinforcements were projected onto the screen from behind by either one or two of three inline projectors located behind the screen. The four values -- one value from each dimension -- necessary to make up a stimulus were projected simultaneously, superimposed on each other. Each projector contained eight transparencies, one for each value (four dimensions times two values per dimension). The lamps for the set of values which made up a particular stimulus were turned on from the control unit. All of these lamps were identical and drew current from a regulated power supply to keep intensities equal.

The subject responded by pressing on one side or the other of a divided response panel which formed the sloping top of a small plastic box. This box measured 6 inches wide by 4 inches deep. The two panels operated Microswitches under them which were connected to the control unit. For the successive presentation group, the left panel was labelled "YES" and the right panel "NO". For the simultaneous presentation groups, both panel halves were blank. The reinforcements were projected on the same screen as the stimuli; directly below the stimulus in the successive presentation conditions, or between and below the stimuli in the simultaneous presentation conditions.

The experimenter was seated in a position behind and to the right of the subject, allowing the experimenter to see both the screen and the subject's response panel. The experimenter's control panel contained indicators which duplicated the subject's responses and reinforcements. The experimenter's controls consisted of pushbuttons to 1) mark the data output for the beginning of a new subject's data, 2) set the control unit to the starting point of the next problem, and 3) start the new problem. In addition, there was a switch to select the appropriate projectors for successive or simultaneous presentation.

The control unit was located in the room adjacent to the experiment room. It consisted of three sections: input, logic, and output. The two switches for the subject's responses, the experimenter's controls, and a punched paper tape reader made up the input. A continuous loop control tape was placed in the reader. On this tape, a special code identified the beginning of each problem. When the logic section sensed this code, information which defined the correct dimension and value of the concept for that problem was read from the tape and stored. Following this information, the tape contained a sequence of codes which defined each stimulus for that problem. This whole pattern was repeated for each problem to be used in the experiment, including the practice problems. Following the last problem, it was only necessary to continue reading the tape to arrive at the beginning for the next subject.

The logic section of the apparatus consisted of a collection of Digital Equipment Corporation K Series solid state logic cards, interconnected to run the experiment. The logic section transformed the paper tape input into the signals necessary to present each stimulus. It accepted the subject's responses and calculated and sent signals to display the reinforcement for each trial. The logic counted the current string of consecutive correct responses up to eight and ended the problem if it was a practice problem; in experimental problems, the logic merely counted eight trials then ended the problem. The logic put the events of each trial in order while timing and spacing them. Finally, it calculated and output the data from the experiment.

The output section of the control unit was a paper tape punch. It recorded one line of data for each trial of a problem. The information recorded on each trial was 1) the correct response, 2) the subject's response, 3) the reinforcement, and 4) three bits of bookkeeping information. Only number 2) above was really necessary, but the inclusion of the rest as redundant information allowed cross-checking of the data and reconstruction of some data which would have otherwise been lost. The physical arrangement of the experiment appears in Figure 1.

Stimuli

The following description applies to the stimuli as they were seen by the subject on the screen during the experiment. In the successive presentation conditions, the stimulus was in the center of the screen; in the simultaneous presentation conditions, the two stimuli were in the middle of the screen vertically and separated horizontally. The left hand stimulus for the simultaneous groups was identical to the stimulus for the successive groups; the right hand stimulus of the simultaneous groups was the complement or opposite ot the other stimulus.

A stimulus was composed of four dimensions, there being one of two possible values present on each dimension. The four dimensions and their values were color: red and blue, bar: horizontal and vertical, shape: circle and square, and diagonal lines: left and right. the color dimension was the two inch square of background color of the stimulus. Eight of the sixteen possible stimuli are exhibited in Figure 2. The color is not indicated, but there were eight stimuli like those in the figure which were red and eight more which were

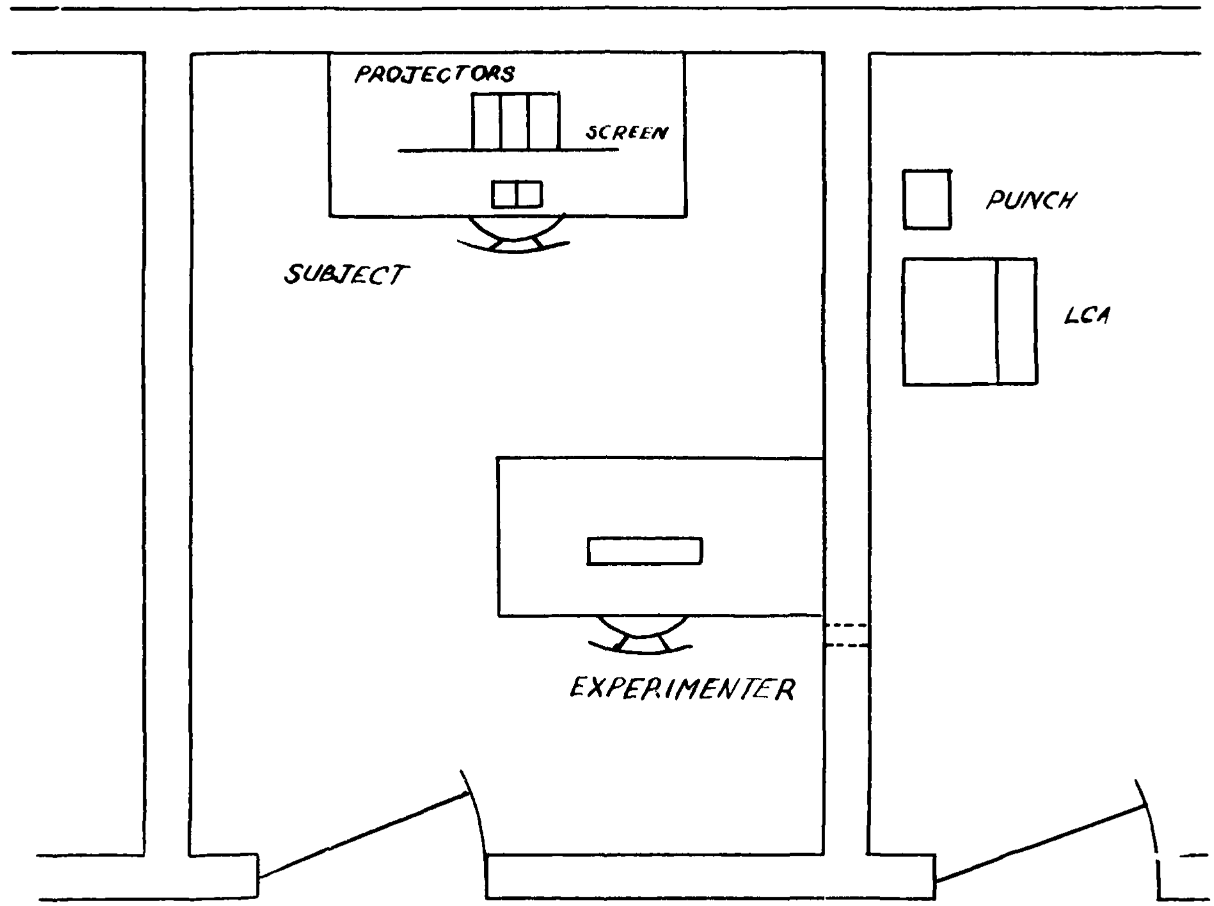

Figure 1. Physical arrangement of the experiment.

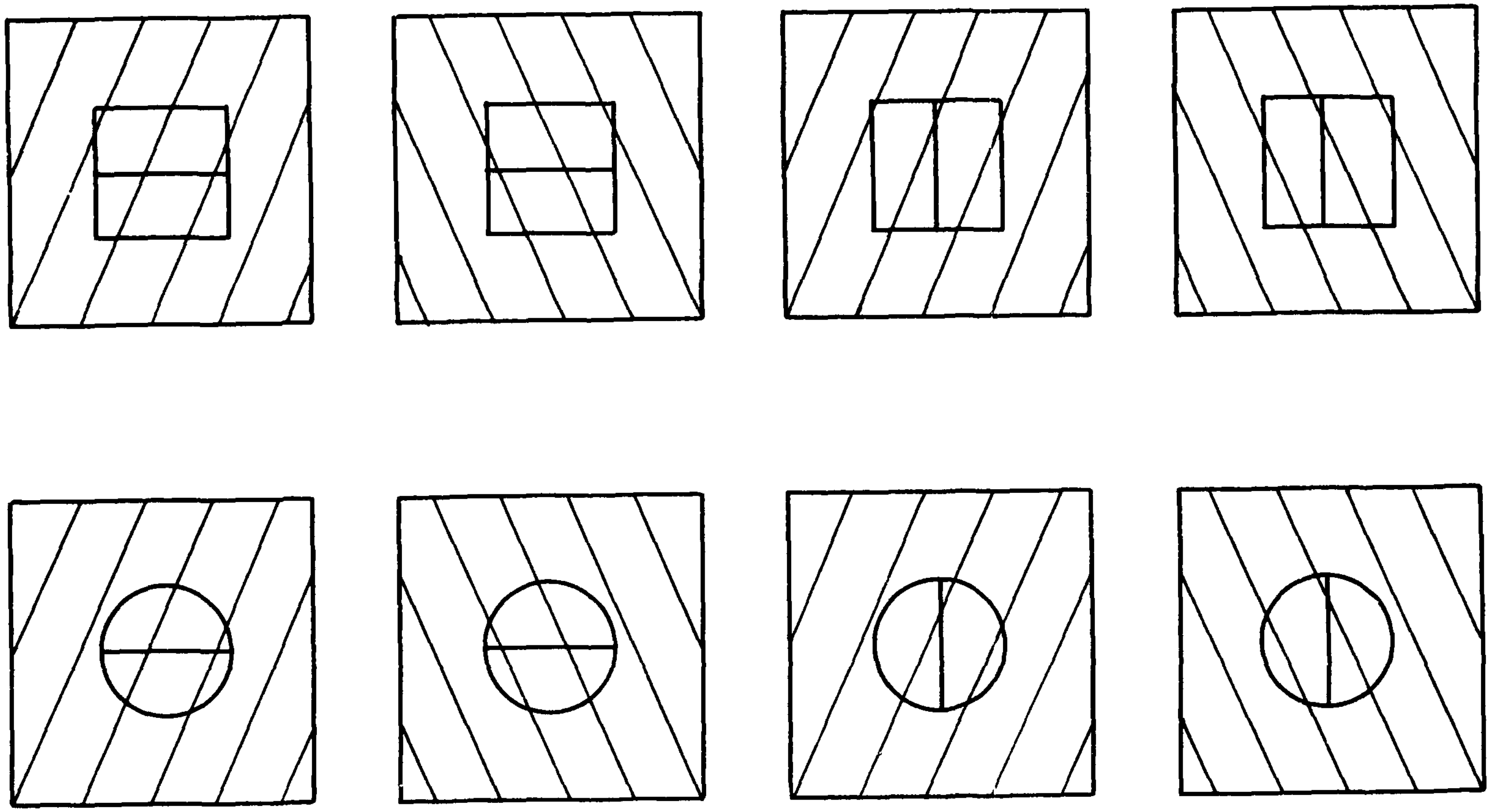

Figure 2. Stimuli. (lines within background are white -- background colored)

blue. Unlike the figure, the actual stimuli were made of broad lines of white light projected over the background of colored light. The diagonal lines were narrower than the lines of the circle, square, or bar. The location of the stimuli for both presentation modes is shown in Figure 3 below, in the section "Conditions". The reinforcement presented to the subject consisted of one of the words "RIGHT" and "WRONG" projected in large white block letters under the stimulus for successive presentation conditions, or under and between the stimuli for simultaneous presentation conditions. The location of the reinforcement is also shown in Figure 3.

Tasks

The correct concepts for the three practice and sixteen experimental problems are presented below in Table 3. Certain restrictions on the sequence of problems were made. Each value of each dimension was the concept twice. No dimension followed itself immediately in either the same or opposite value; in other words, there were no "reversal shifts" between problems. These restrictions were not apparent to the subject since he was told before each problem that any of the possible concepts might be correct each time. In debriefing the subjects, it was found that they actually were not aware of any restrictions.

Orders

Stimulus orders were prepared observing certain limitations. In making up a problem, an order of correct answers was assigned to the relevant dimension. For successive presentation conditions this amounted to making the stimulus either an example or a non-example of the concept -- a "yes" or a "no". In the simultaneous presentation conditions, the left hand stimulus was the same as the successive presentation stimulus for the same trial. In this mode, the procedure was to make either the left or right stimulus the example of the concept.

There were four different sequences used. The values of the relevant dimension of order III were the complements of the relevant dimension values of order I. Similarly, IV was the complement of II on the relevant dimension. Values were assigned to the three irrelevant dimensions in such a way that each of the sixteen possible stimuli was used the same number of times. The four orders are shown in Table 2.

These four orders were then assigned to the sixteen problems described above in the section "Tasks". Each order was assigned to two problems as shown in Table 2, and in its complementary form to two other problems. In other words, order I defines the correct answers to be Y, Y, N, Y, ... This sequence was used for two problems; on two others, order I was complemented -- each 0 became a 1, each 1 a 0 -- so that the correct answers were N, N, Y, N, ... The same order was never used with two consecutive problems.

Since the performance of the individual was the prime interest of the experiment, the same sequence of problems and order of stimuli within problems were used for all subjects; no counterbalancing was undertaken.

Table 2. Orders of stimuli.

Order	I				II				III				IV			
Dimension	R	I1	I2	I3	R	I1	I2	I3	R	I1	I2	I3	R	I1	I2	I3
Trial 1	1	1	1	1	0	0	0	0	0	1	0	0	1	0	0	1
2	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1	0
3	0	1	1	0	0	1	0	1	1	1	0	0	1	0	0	0
4	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	1
5	0	0	1	1	1	1	0	1	1	1	1	1	0	1	0	0
6	0	0	0	1	1	0	1	0	1	1	0	1	0	0	0	0
7	0	1	0	1	0	1	1	0	1	0	0	1	1	0	1	1
8	1	0	1	0	0	0	1	1	0	1	1	1	1	1	1	0

R = relevant dimension Ii = irrelevant dimension

Table 3. Correct concepts and orders used.

Problem	Dimension	Value	Order	
1	Color	Blue	I	C
2	Diagonals	Left	IV	
3	Shape	Circle	III	
4	Diagonals	Right	I	C
5	Bar	Horiz	III	
6	Shape	Square	II	C
7	Bar	Horiz	IV	
8	Color	Blue	II	C
9	Diagonals	Right	IV	C
10	Bar	Vert	I	C
11	Shape	Square	IV	C
12	Color	Red	III	
13	Shape	Circle	II	
14	Color	Red	I	
15	Diagonals	Left	II	
16	Bar	Vert	III	C
P1	Bar	Horiz	-	
P2	Color	Red	-	
P3	Shape	Circle	-	

C = complemented

Conditions

The instruction content and mode of stimulus presentation were the two experimental variables used in this experiment. Two values of each variable were combined to make four groups.

The intent of the instructions was to give a minimally sufficient understanding of the task to the subject. In the short (S) instruction groups, no information was given which would suggest a strategy for the task; the instructions just defined the necessary rules of the task. The long (L) instructions however, suggested the need for initial guessing and the idea of eliminating possibilities from some larger set of possibilities. Although there had to be some differences in instructions between the two groups of each level of instructions, the two sets of instructions were made as parallel as possible.

The two values of the presentation variable allowed for one or two stimuli to be seen at one time. Together with the instruction variable, this made four experimental groups: 1S, 1L, 2S, and 2L. The nature of the successive presentation groups was that a single stimulus was presented on each trial. The simultaneous groups saw a double stimulus on each trial. The left hand stimulus for the simultaneous groups was identical to the successive group stimulus for the same problem and trial. The right hand stimulus for the simultaneous groups was the complement of the left hand stimulus: each dimension on the right showed the opposite value from that dimension on the left. Figure 3 below shows a sample trial as it would appear to both the successive and simultaneous presentation groups.

Instructions

Following are the instructions for each group as they were read to and by the subject. The first set is for the successive presentation, long instruction (1L) group:

> This is a concept identification problem. You will see some designs on the screen in front of you. Each design has four characteristics: a color, a shape, a bar and shading. The color will be red or blue, the shape circle or square, the bar will be vertical or horizontal, and the shading will slope to the right or to the left. A given concept will depend on one and only one of the characteristics, so that each design either shows or does not show that concept. For example, if the concept is RED, you would press "YES" if the design is red or "NO" if the design is blue. It makes no difference in this example if the shape is circle or square, if the bar is horizontal or vertical, or if the shading is to the right or to the left; the only thing that makes any difference is the color. After you press the button, the screen will say "RIGHT" or "WRONG", depending on your answer.
>
> The problem is to find out what the concept is by looking at the designs, making your answer, and finding out if you were right. For each problem, the concept is the same until the end of that problem. Also, on each problem any one of the possible concepts might be the correct one. The order in which the designs appear makes no difference. You will have three practice problems first. On the first trial of each problem, you will have no idea what the concept might be, so all you can do is guess your answer. But when you find out about the answer, you will have some information about the correct concept. We will try the practice problems now.

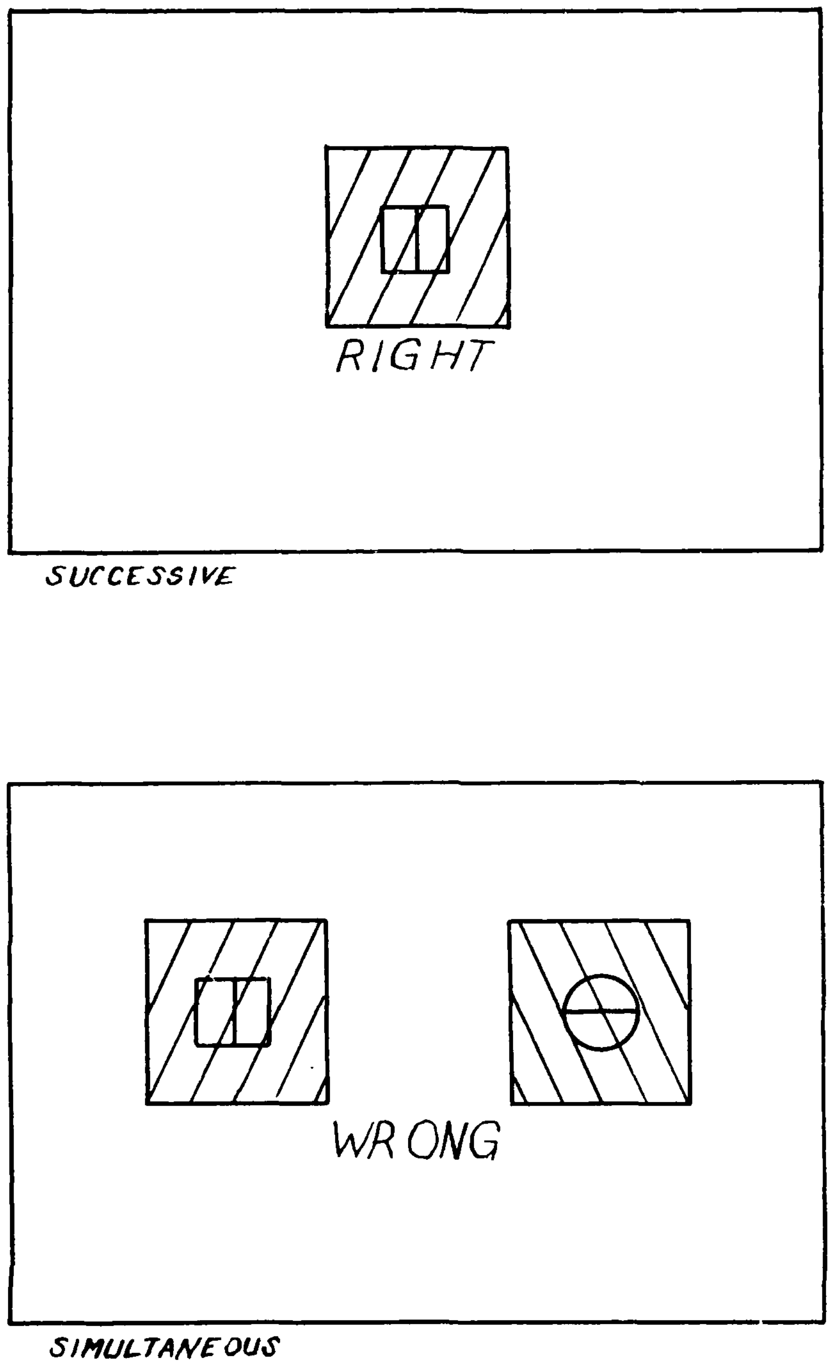

Figure 3. Stimulus and reinforcement location.

These instructions are for the successive presentation, short instruction (1S) group:

This is a concept identification problem. You will see some designs on the screen in front of you. A given concept will depend on one and only one thing in the designs, so that each design either shows or does not show that concept. You will see a design on the screen, then press "YES" if you think the design shows the concept or "NO" if you think the design does not show the concept. After you press the button, the screen will say "RIGHT" or "WRONG", depending on your answer.

The problem is to find out what the concept is by looking at the designs, making your answer, and finding out if you were right. For each problem, the concept is the same until the end of that problem. Also, on each problem any one of the possible concepts might be the correct one. The order in which the designs appear makes no difference. You will have three practice problems first. On the first trial of each problem, all you can do is guess your answer. We will try the practice problems now.

This set of instructions applied to the simultaneous presentation, long instruction (2L) group:

This is a concept identification problem. You will see some designs on the screen in front of you. Each design has four characteristics: a color, a shape, a bar and shading. The color will be red or green, the shape circle or square, the bar will be vertical or horizontal, and the shading will slope to the right or to the left. A given concept will depend on one and only one of the characteristics, so that each design either shows or does not show that concept. Only one of the designs on the screen will show the concept. For example, if the concept is RED, you would press the button by the red design. It makes no difference in this example if the shape is circle or square, if the bar is horizontal or vertical, or if the shading is to the right or to the left; the only thing that makes any difference is the color. After you press the button, the screen will say "RIGHT" or "WRONG", depending on your answer.

The problem is to find out what the concept is by looking at the designs, making your answer, and finding out if you were right. For each problem, the concept is the same until the end of that problem. Also, on each problem any one of the possible concepts might be the correct one. Neither the order in which the designs appear nor the side which they appear on makes any difference. You will have three practice problems first. On the first trial of each problem, you will have no idea what the concept might be, so all you can do is guess your answer. But when you find out about the answer, you will have some information about the correct concept. We will try the practice problems now.

And finally, the instructions for the simultaneous presentation, short instruction (2S) group:

This is a concept identification problem. You will see some designs on the screen in front of you. A given concept will depend on one and only one thing in the designs, so that each design either shows or does not show that concept. You will see a pair of designs on the screen, only one of which shows the concept, then press the button by the design you think shows the concept. After you press the button, the screen will say "RIGHT" or "WRONG", depending on your answer.

The problem is to find out what the concept is by looking at the designs, making your answer, and finding out if you were right. For each problem, the concept is the same until the end of that problem. Also, on each problem any one of the possible concepts might be the correct one. Neither the order in which the designs appear nor the side which they appear on makes any difference. You will have three practice problems first. On the first trial of each problem, all you can do is guess your answer. We will try the practice problems now.

Procedure

The subjects performed the task of this experiment individually. When they entered the experiment room, the door was closed and the experimenter introduced himself. The subject was seated in front of the stimulus display and response panels. Bookkeeping matters were recorded, after which the experimenter handed the subject a copy of the instructions, directing him to read along as the experimenter read them aloud.

After the instructions, the subject had an opportunity to ask questions. If he had any, the experimenter tried to answer them by rereading or paraphrasing the initial instructions with appropriate emphasis; this was done with care not to change the intent of the instructions. Many questions were deferred to be answered by experience in the practice problems.

During the practice problems, the experimenter answered the subject's questions, making sure he had learned the task by the end of the last practice problem. A practice problem was ended by the subject producing a string of eight consecutive correct responses. Before each practice and experimental problem, the subject was reminded that one and only one simple concept would be correct, and that it could be any one of the possible concepts. The chain of events in a problem was as follows.

1. The experimenter turned on the first stimulus.
2. The subject had unlimited time to respond.
3. Immediately following the subject's response, the reinforcement came on with the stimulus remaining present.
4. Following a fixed interval of approximately three seconds, the stimulus and reinforcement went off.

5. Within about one half second, the next stimulus appeared, unless the problem was already completed.

Steps 2 through 5 above were repeated until eight consecutive "RIGHT" reinforcements were made during practice problems, or until eight trials had been completed in experimental problems.

When the problem was finished, the subject was asked to name the concept. If his answer was incorrect, the experimenter informed the subject of the correct concept. If the subject responded with an illegal concept such as a compound or complex concept, or one outside the intended stimulus space, the experimenter would provide the correct concept and the necessary information to redefine the task correctly.

After the three practice problems, the experimenter explained that the experimental problems would be the same, except that the subject would have only "a limited number of trials" in which to solve the problems. Following the eighth or middle problem, the subject was offered a short break. If he took it, he was allowed about a minute in the experimental room during which the experimenter refrained from discussing the problems. After the break, the second set of eight problems was undertaken.

At the conclusion of the experiment, the experimenter explained the purposes and techniques of the experiment and answered any questions the subject had. The whole session was completed in less than twenty-five minutes.

ANALYSIS

Definition of the Bower and Trabasso Model

The Bower and Trabasso (Atkinson, et.al., 1965) model for concept identification can be applied to the experiment described in this paper. The model assumes the subject to be in a guessing state at the start of a problem. The subject has no hypothesis about the correct concept at the beginning of the problem. In the guessing state, the subject guesses wrong (makes an error) with a probability of p. It is assumed here that $p = 1/2$. When an error occurs, the subject selects an hypothesis to replace any previously selected hypotheses. This hypothesis is consistent with the information available on the trial of the error. The probability of selecting the correct hypothesis is c . The subject is assumed to retain this newly selected hypothesis and respond according to it until he makes an error, at which time the same hypothesis selection procedure is again performed. Since the experiment described above provided for only a fixed number of trials, it is also assumed that if a subject is still in the guessing state at the end of a problem -- no errors have occurred -- he will select an hypothesis when asked for the concept just as if he had made an error on the last trial. In addition, he will report his current hypothesis at the end of a problem without resampling, unless he has made an error on the final trial, in which case the subject resamples as for any error trial.

Stimulus Information

In the past, the Bower and Trabasso model has not involved any record of or information about the actual stimuli to which the subject was responding. It was assumed that the stimuli were sufficiently random and numerous that the probability of an error given an incorrect hypothesis was equal to p or 1/2. In the present analysis the actual stimulus sequence is employed to give a more detailed account of the protocol. The stimulus sequence may show that some of the irrelevant dimensions were not hypotheses which would produce the subject's actual responses.

Consider the following protocol where a "1" indicates an error and a "0" a correct response. Also consider the stimulus sequence shown where a "1" is one value of the given dimension and a "0" the other value, i.e., 1 = red and 0 = blue for the dimension color.

Trial	1	2	3	4	5	6	7	8	
Protocol	0	1	1	0	0	0	1	0	(unsolved)
Dimension 1	0	0	1	0	1	1	1	0	
Dimension 2	0	1	0	1	1	1	0	1	
Dimension 3	0	0	0	1	0	1	1	0	
Dimension 4	0	0	1	1	0	0	0	1	

The relevant dimension in the above sequence is dimension 1. The subject made an error on trial 2 and also made further errors, indicating that he had selected an irrelevant dimension as his hypothesis after the error on trial 2. On trial 2, the correct answer -- the value of the relevant dimension -- was a "0" or "no". Since the subject made an error on trial 3 and the answer for that trial was "1" or "yes", his answer had to have been "0". Dimension 3 is the only hypothesis which could have led him to make that answer, since it is the only dimension for which the value does not change from trial 2 to trial 3 while the correct dimension does change values. Now consider the error on trial 3 and the succeeding portion of the protocol through trial 7. In order for the observed protocol to have occurred, the Bower and Trabasso model says that the subject must have selected an hypothesis which was perfectly correlated with the relevant dimension from trials 3 through 6 inclusive, and then changed values between trials 6 and 7 where the relevant dimension did not change values. This is the only way the protocol could be generated using the assumption of the model. By examining the protocol and the stimulus sequence, it can be seen that none of the irrelevant dimensions are consistent with the protocol from trial 3 through trial 7. In other words, it is impossible for this subject to have used the Bower and Trabasso strategy in this problem. The probability of the model given the data is zero.

In some cases then, very definite statements about the model can be made by using the stimulus information in the analysis of the data. In general it would seem that much better evaluation can be obtained with more available information being used.

Probability of a Protocol

Previous tests of the Bower and Trabasso model have run subjects to a criterion of consecutive correct responses, allowing the evaluation to assume that the final error led to selection of the correct hypothesis. The expression for the probability of a protocol under the previous evaluation was

$$P(D|M) = (1/2)^{t} (1-c)^{k-1} c$$

where: t is the trial of last error
k is the number of errors and
c is the probability of selecting the correct hypothesis following an error.

D is the observed data -- the protocol -- and M represents the model assumptions. The probability of any response being an error up to and including the final error is 1/2, when p is assumed to be 1/2. On all but one of the k errors, an incorrect hypothesis was selected with probability 1-c, and the correct hypothesis was selected with probability c following the final error.

If the stimulus information is considered, an expression for the probability of a protocol may still be derived from the model. Up to and including the first error, the subject is guessing without an hypothesis and makes each response with probability of an error of 1/2. When the first error occurs, the subject selects an hypothesis. If it is correct, the probability of the event is c, giving in this case

$$P(D|M) = (1/2)^{t'} c$$

where: t' is the trial of the first error.

If the subject makes no errors throughout the eight trials and gives the correct concept when asked for it at the end of the problem, it is necessary to assume that he was in the guessing state for eight trials ($t' = 8$) and selected the correct hypothesis when asked; the probability of this last event would be c . In this case, then

$$P(D|M) = (1/2)^{8} c \; ;$$

however, if the subject gives the wrong concept after eight errorless trials, the probability expression is

$$P(D|M) = (1/2)^{8} (1-c) \; .$$

If the subject has made an error and selects an incorrect hypothesis as indicated by further errors, the expression becomes more involved. With probability $1-c$ he selects one of the three incorrect hypotheses. To preserve the mathematical tractability of this analysis, let the three incorrect hypotheses each have a probability of being selected of $(1/3)(1-c)$. This assumption also prevents the model from becoming a four- rather than a one-parameter model. As shown in the section "Stimulus Information", it can be determined whether each incorrect hypothesis could produce the observed protocol. These consistent hypotheses can then be counted. For the case of a single error in an unsolved problem, the probability becomes

$$P(D|M) = (1/2)^{t'} (h/3)(1-c)$$

where: h is the number of consistent hypotheses.

The probability of the part of the protocol up to and including the error is $(1/2)^{t'}$ and the probability of the rest of the protocol is $(h/3)(1-c)$. Note that it is not necessary to consider each response after the error to occur with a probability of 1/2 since the $(h/3)(1-c)$ is the probability of the whole sequence. To extend the expression's generality to all possible protocols, it is necessary to consider both solved and unsolved problems with any number of errors. This derivation from the model gives

$$P(D|M) = (1/2)^{t'} [\Pi(h_i/3)] (1-c)^{k-s} c^{s}$$

where: s = 1 if solved,
s = 0 if not solved,
k is the number of errors.

Notice that there are either k or k-1 terms of the form $(h/3)(1-c)$ -- one for each non-final error. There are k if the problem is unsolved (s = 0) or k-1 if the problem is solved (s = 1). The $\Pi(h/3)$ term has one $h_i/3$ for each error which did not lead to the correct concept. Note also that if the sequence of responses between any two errors is inconsistent with all three irrelevant dimensions, the h for that term is zero, which makes $\Pi(h_i/3)$ zero, and therefore $P(D|M) = 0$.

Sequential Application of Bayes' Theorem

Bayes' Theorem states that

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)} .$$

$P(A)$ is the prior probability of A, and $P(A|B)$ is the probability of A after observing B.

Let D_i be the data (protocol) observed on a given problem and let $P(M)$ be the initial "subjective" probability given to the model's occurrence. Substituting into Bayes' Theorem,

$$P(M|D_i) = \frac{P(D_i|M)\ P(M)}{P(D_i)} .$$

The discussion above in "Probability of a Protocol" shows that the model defines a value for $P(D_i|M)$. $P(M)$ can be more or less arbitrary since it is the subjective prior probability. $P(D_i)$ is the unconditional probability of a given protocol. $P(D_i)$ can also be expressed as shown below (Parzen, 1960, p.119).

$$\sum_j P(D_i|M_j)\ P(M_j) \ ;$$

or the sum of the probabilities of the protocol under all possible models. Since the set of all M_j cannot be defined for calculating this sum, it is necessary to estimate $P(D_i)$. The best estimate available is the observed distribution of protocols from all of the subjects involved in the experiment. This distribution has to be on the basis of each problem because of the different stimulus sequences used for different problems. This set of protocols is a sample of protocols from subjects who could be operating under any of the possible models in the set. Making these substitutions into Bayes' Theorem, a posterior probability of the model given the data can be calculated.

Since each subject performed sixteen problems with each dimension being relevant four times, it is possible to make an even more complete evaluation by handling together the groups of four problems which have the same relevant dimensions. The parameter is assumed to be constant over the four problems.

Let the probability calculated from the protocol be used as the prior probability for the next protocol. This substitution may be applied sequentially throughout all of the problems to be analyzed. Given this sequential application, the probability of the model given the data should converge on the same value regardless of the initial prior probability if $0 < p < 1$ (Blackwell and Dubins, 1962).

$$P(M|D_1) = \frac{P(D_1|M)\ P(M)}{P(D_1)}$$

$$P(M|D_2) = \frac{P(D_2|M)\ P'(M)}{P(D_2)} = \frac{P(D_2|M)\ P(M|D_1)}{P(D_2)}$$

$$P(M|D_2) = \frac{P(D_2|M)\ \dfrac{P(D_1|M)\ P(M)}{P(D_1)}}{P(D_2)}$$

$$P(M|D_2) = \frac{P(D_2|M)\ P(D_1|M)\ P(M)}{P(D_2)\ P(D_1)}$$

$$P(M|D.) = \left[\prod_{i=1}^{n} \frac{P(D_i|M)}{P(D_i)}\right] P(M)$$

where: D. represents all the data and
$n = 4$ in this case.

This sequential application of Bayes' Theorem then gives a well defined expression for the probability of the given model for one subject, on one dimension.

By substituting in the expression for $P(M|D_i)$ that was derived above, the following well defined expression for the probability of the model given the data is obtained.

$$P(M|D.) = \frac{(1/2)^{\Sigma t'}\ \dfrac{\Pi(\Pi h)}{3^{\Sigma(k-s)}}\ (1-c)^{\Sigma(k-s)}\ c^{\Sigma s}\ P(M)}{\Pi P(D_i)}$$

Estimate of the Model Parameter

According to the definition of the Bower and Trabasso model, it is a one parameter model when p is assumed to be 1/2 . The parameter is c, the probability of selecting the correct hypothesis following an error. It is possible to obtain a maximum likelihood estimate from the expression derived above. This is done by maximizing $P(M|D.)$ with respect to c. The value of $\hat{c}$ is found by taking the derivative with respect to c of $P(M|D.)$, setting it equal to zero, and solving the resulting equation for c ($\hat{c}$).

Since only terms involving c affect the derivative, $P(M|D.)$ may be simplified to

$$P(M|D.) = K\ (1-c)^{\Sigma(k-s)}\ c^{\Sigma s}$$

where: K represents all terms not involving c.

Taking the derivative with respect to c,

$$D_c[\ P(M|D.)\ ] = D_c[\ K\ (1-c)^{\Sigma(k-s)}\ c^{\Sigma s}\ ]$$

$$= K[\ (1-c)^{\Sigma(k-s)}\ \Sigma s\ c^{(\Sigma s)-1} + c^{\Sigma s}\ \Sigma(k-s)\ (1-c)^{\Sigma(k-s)-1}\ (-1)\ ]\ .$$

Setting this equal to zero and solving for c ($\hat{c}$),

$$\hat{c} = \frac{\Sigma s}{\Sigma s + \Sigma(k-s)} = \frac{\Sigma s}{\Sigma k}\ .$$

The denominator of the expression is then the total number of errors produced by all the problems involved in the evaluation. The numerator is the number of problems solved. In the special case that all the problems are solved, this estimate of c is identical to the estimate derived by Bower and Trabasso (Atkinson, et. al., 1965, p. 71) for their group data analysis. When there are unsolved problems however, the estimate has the following property. It is not equal to the average of the individual estimates of c for each problem, where the estimate of c is zero for an unsolved problem. For a small number of errors in the unsolved problems, the estimate tends to be larger than the average of individual estimates; it tends to be smaller than the average when a large number of errors occur in unsolved problems.

Evaluation Characteristics

Given the expression derived for $P(M|D.)$, several characteristics of the evaluation technique may be noted. Consider the final form of the expression below and note that there are three factors.

$$P(M|D.) = \left[\prod \frac{P(D_i|M)}{P(D_i)} \right] \cdot P(M) = \frac{P(D.|M)}{P(D.)} \cdot P(M) .$$

The first factor -- $P(M)$ -- is the initial evaluation of the model. This is constant and is the starting point of the evaluation. The second factor -- $P(D.)$ -- is the unconditional probability of the data observed. This can be considered the part of the expression which refers to the actual universe; it is a description of the way things actually exist. The important factor in the expression is the third factor -- $P(D.|M)$ -- which is the predictive element of the model. There are interesting comparisons to be made between $P(D.)$ and $P(D.|M)$.

If $P(D.|M)$ is less than $P(D.)$, then the model does not add anything to the predictive power of the descriptive $P(D.)$. One would predict better for a given subject by using the distribution of data from previous like problems of many subjects. Looking at the whole equation, when $P(D.|M)$ is less than $P(D.)$, $P(M|D.)$ is less than $P(M)$. In other words, this situation says that it is less likely that the model is true after we have observed some data than it was before. In the other case, when $P(D.|M)$ is greater than $P(D.)$, an increase in the probability of the model occurs from $P(M)$ to $P(M|D.)$. An index of the increase is then the ratio of $P(D.|M)$ to $P(D.)$. If the ratio is less than one, the model adds no information and should be rejected. If the ratio is greater than one, the model is predictive.

The result of this technique is the probability that the model generated the observed data. A decision function for accepting the model as tenable would depend on two things. The first is that the ratio of $P(D.|M)$ to $P(M)$ be greater than one. Secondly, the final probability of the model depends on the initial probability assigned to the model. For a sufficiently large $P(M)$, a ratio only slightly greater than one will generate a $P(M|D.)$ equal to one. Below is the formal statement of the evaluative function.

$$P(M|D.) = R \cdot P(M) \qquad \text{for } R \leq P(M)^{-1}$$
$$= 1 \qquad \text{for } R > P(M)^{-1}$$
$$\text{where:} \quad R = \frac{P(D.|M)}{P(D.)} \quad .$$

$P(M|D.)$ could be defined as zero for R less than one since no information is added as discussed above, although this is not a mathematical conclusion drawn from the function itself.

Recall that a separate probability is calculated for each subject for each dimension, with four problems per dimension. If the subject is assumed to be stable with respect to his strategy throughout all sixteen problems, none of these four probabilities may be zero without completely eliminating any possibility of the subject operating under the model. More specifically, if one or more problems is an impossible protocol under the model, then $P(M|D.)$ for that subject is zero.

RESULTS

Effects of General Conditions

The concept identification task of the present experiment showed two characteristics. The mode of presentation affected all of the general measures used. In addition there was very definitely no stability of performance over the time involved in the experiment.

The general measures analysed in this section were the probability of solution, the number of errors, and the trial of the last error for solved porblems. These three measures are summarized below in Tables 4, 5, and 6. In these tables, if no probability is given for the F statistic, the probability was greater than .01.

Figure 4 below shows the probability of solution over problems. There is definitely an increase in probability; the subjects appear to be approaching an asymptote of one by the last problem. Figure 5 shows the differences in number of solved problems between the successive presentation groups (1S and 1L) and the simultaneous presentation groups (2S and 2L). Figures 6 and 7 indicate the effects of problem number and successive versus simultaneous presentation on the number of errors.

Although problem number is a significant factor in trial of last error, Figure 8 indicates that the differences are between individual problem numbers due to the different stimulus orders and do not indicate a monotonic decrease. Figure 9 indicates the varying difficulties of each of the four dimensions on each of its occurrences and for the average of all four occurrences.

Table 4. Solution probability analysis of variance.

Variable	SS	df	MS	F	p
Presentation	3.4789	1	3.4789	22.72	<.001
Instructions	.0137	1	.0137	.09	
Problem number	16.4285	15	1.0952	7.15	<.001
Pres x Inst	.0595	1	.0595	.39	
Pres x Prob	2.5188	15	.1679	1.10	
Inst x Prob	2.5883	15	.1726	1.13	
Pres x Inst x Prob	1.5130	15	.1009	.66	
Error	171.5184	1120	.1531		
Total	198.1191	1183			

Table 5. Number of errors analysis of variance.

Variable	SS	df	MS	F	p
Presentation	21.9444	1	21.9444	9.74	.002
Instructions	8.6219	1	8.6219	3.83	
Problem number	193.6609	15	12.9107	5.73	<.001
Pres x Inst	.7770	1	.7770	.34	
Pres x Prob	73.4727	15	4.8982	2.17	.006
Inst x Prob	24.3058	15	1.6209	.72	
Pres x Inst x Prob	39.8188	15	2.6546	1.18	
Error	2522.3770	1120	2.2521		
Total	2884.9785	1183			

Table 6. Last error analysis of variance, solved problems.

Variable	SS	df	MS	F	p
Presentation	14.0356	1	14.0356	3.19	
Instructions	4.5087	1	4.5087	1.03	
Problem number	277.2831	15	18.4855	4.20	<.001
Pres x Inst	.9106	1	.9106	.21	
Pres x Prob	34.1658	15	2.2777	.52	
Inst x Prob	31.6388	15	2.1093	.48	
Pres x Inst x Prob	124.3761	15	8.2917	1.89	
Error	3822.0782	869	4.3982		
Total	4289.7771	932			

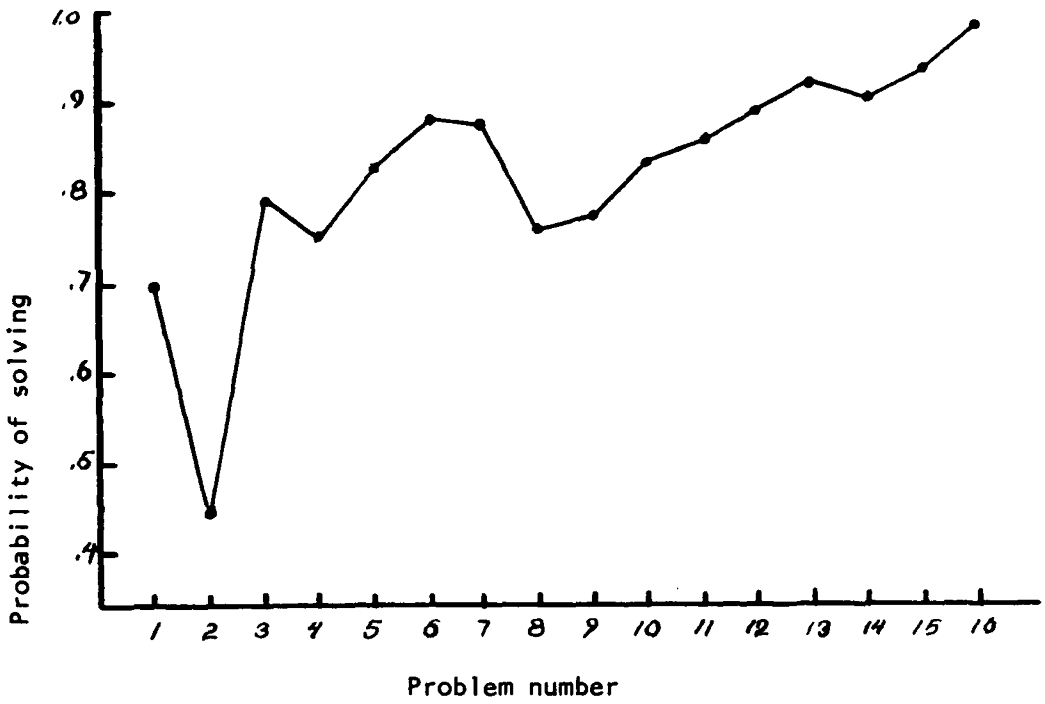

Figure 4. Solution probability

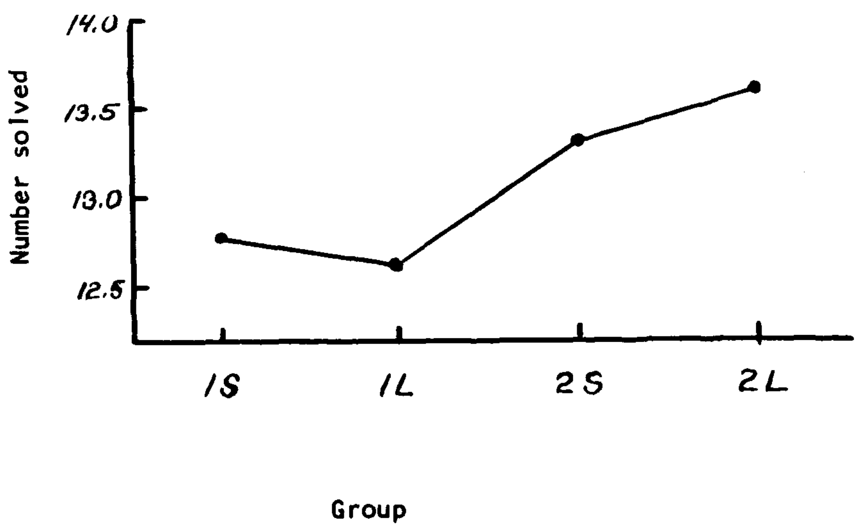

Figure 5. Number solved

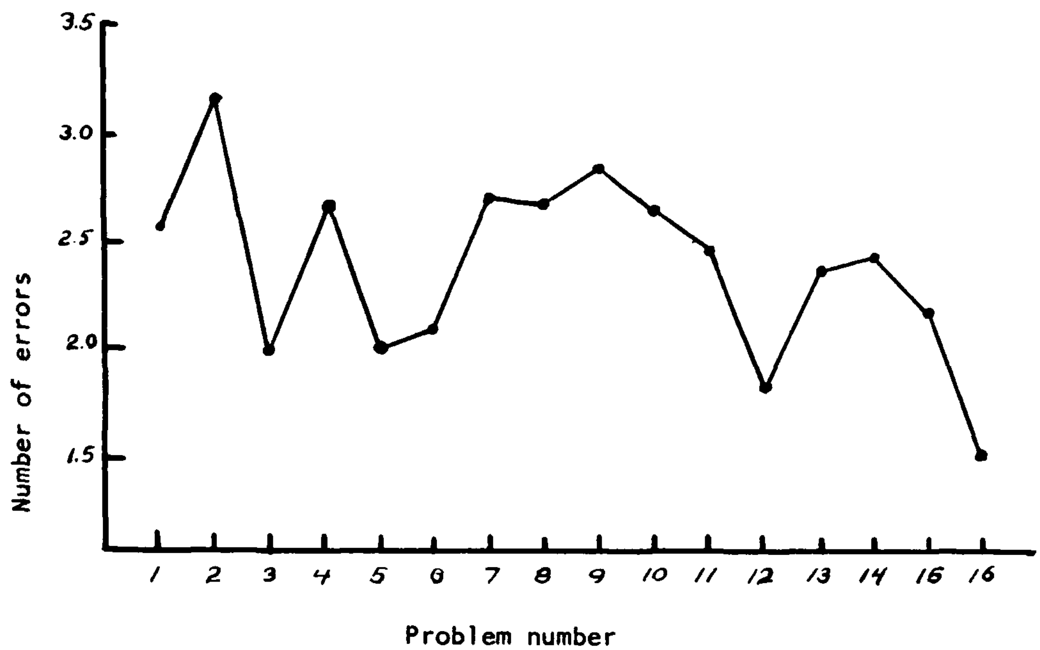

Figure 6. Number of errors.

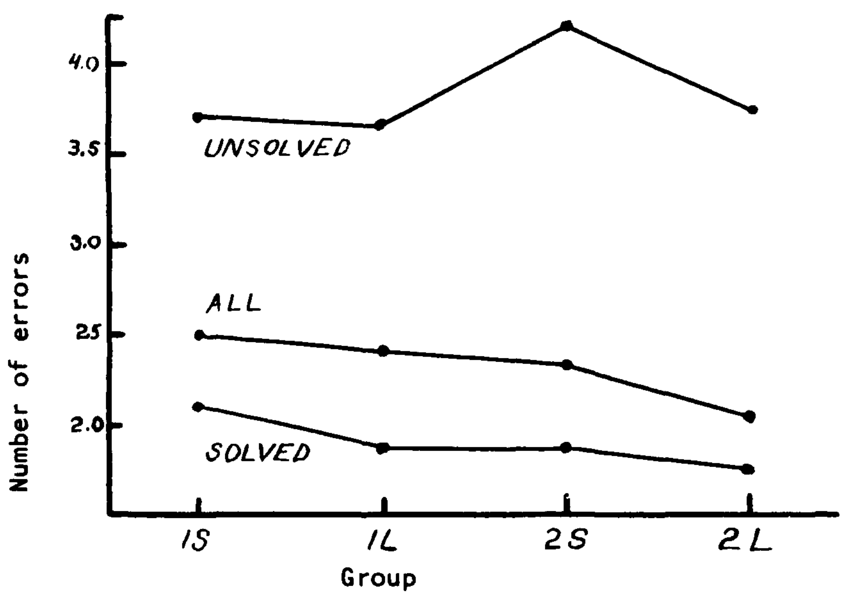

Figure 7. Number of errors by groups.

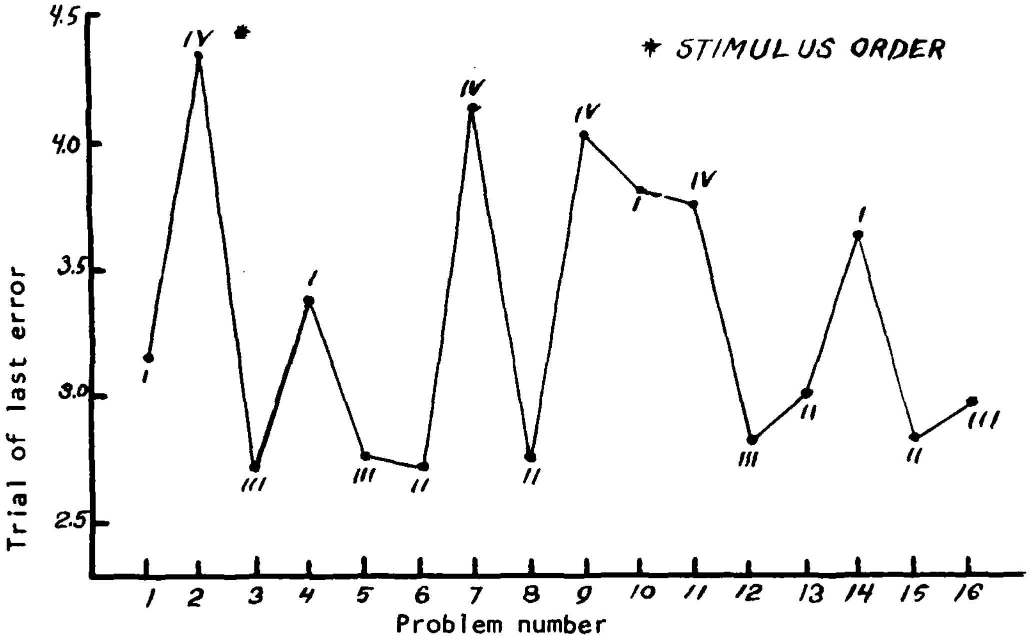

Figure 8. Trial of last error, solved problems.

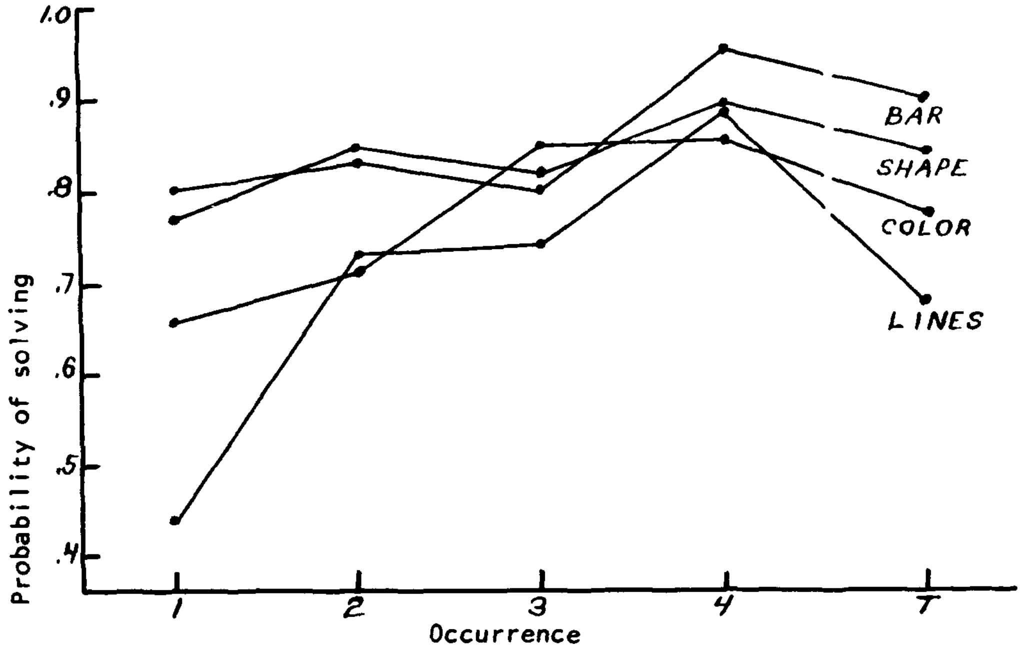

Figure 9. Solution probability by dimension and occurrence.

Figure 10 shows the traditional learning curves for each dimension broken down by successive or simultaneous stimulus presentation. The learning curves appear to be reasonably typical for concept identification experiments.

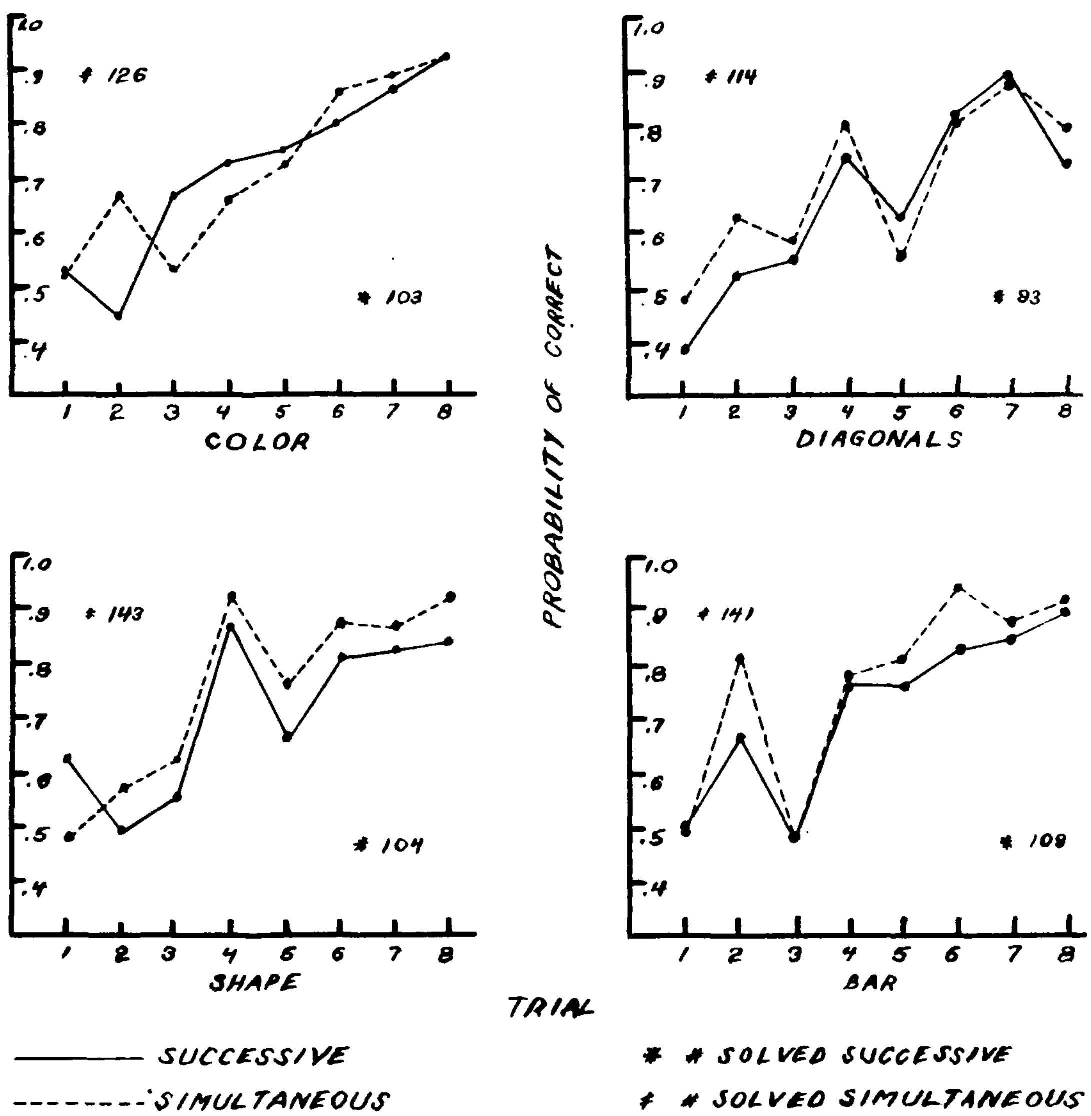

Figure 10. Learning curves.

Bayesian Evaluation

In applying the evaluation technique described above, some very strong statements about the fit of the Bower and Trabasso model to the present data were generated. Table 7 gives the number and percent of subjects from each group with nonzero probabilities of using a Bower and Trabasso strategy. These subjects were determined by selecting for ratios of $P(D.|M)$ to $P(D.)$ which were not zero on any of the four dimensions. These subjects' ratios are listed in Table 8. In effect, Table 8 comprises the results of the experiment. Table 9 lists the estimate of the parameter c for each dimension for the subjects discussed regarding the two previous tables.

Table 7. Possible Bower and Trabasso subjects.

Group	N	Number of Possibles	Percent of Possibles
1 S	16	2	12.5
1 L	19	0	0.0
2 S	19	3	15.8
2 L	20	6	30.0
Total	74	11	14.9

Table 8. Probability ratios.

Subject	Group	Dimension 1	2	3	4
31	1 S	.000+	.001	.000+	.000+
83	1 S	.001	.000+	.000+	.002
15	2 S	.000+	.003	.003	.000+
38	2 S	.000+	.000+	.217	.001
54	2 S	.000+	.004	.000+	.101
3	2 L	.000+	.022	.007	.002
25	2 L	.092	.000+	.925	5.951
*67	2 L	3.627	.033	.027	4.936
71	2 L	.198	.000+	.000+	.007
77	2 L	2.824	.001	.002	.303
86	2 L	.056	.009	.000+	.136

Table 9. Parameter estimates.

Subject	Group	Dimension 1	2	3	4
31	1 S	.33	.33	.38	.18
83	1 S	.27	.44	.44	.19
15	2 S	.25	.57	.33	.33
38	2 S	.67	.31	.57	.21
54	2 S	.38	.43	.57	.27
3	2 L	.80	.18	.67	.27
25	2 L	.80	.57	.80	.67
*67	2 L	.67	.67	.57	.80
71	2 L	.80	.50	.44	.21
77	2 L	.50	.57	.23	.38
86	2 L	.57	.67	.43	.27

DISCUSSION

The level of task complexity involved in this experiment was not great, allowing generally good performance by the subjects. This caused most of the data to fall in the high performance range. The instructions presented to the subjects were adequate at both levels to define the experimental task. Since the intent of the practice problems was to insure that the subjects were past the acquisition phase of the bahavior, it is very possible that the practice problems also leveled out any initial instruction differences. However, on an informal observational level, the long instruction subjects seemed more confident during the practice problems.

The large differences in performance between successive and simultaneous presentation groups are a good indicator that the form of the perceptual object is an important factor in the concept identification process. Efficiency of information processing seems to be much greater when both the stimulus and its complement are present. There are at least two explanations of this result which seem tenable. The processes might involve the storage of information about each value of a dimension independently. Then, elimination of one value of a dimension as being the correct concept would not affect the status of the other value. With simultaneous presentation, both values of each dimension (one in each stimulus) are present for elimination simultaneously. This allows the dimension to be eliminated completely on one trial. In successive presentation, only one value of each dimension is present, so a dimension can only be partially eliminated on a single trial. At least two trials would be required to completely eliminate a dimension. A second possible explanation is that elimination of dimensions operates only on positive instances of the concept. Simultaneous presentation would always present a positive instance, whereas successive presentation would require the subject to perform the extra processing necessary to complement each value of a negative instance in order to operate on it.

Although there is nothing in the definition of the Bower and Trabasso model to indicate that there should be a presentation effect, there was a difference in the number of subjects and their probabilities of using Bower and Trabasso strategies between the successive and simultaneous presentation groups. The simultaneous presentation, long instruction group was the only one which had a reasonably large percentage of subjects who could have even possibly been operating under the Bower and Trabasso model. Simultaneous presentation may show its effect at the selection of a new hypothesis following an error. The selection of an hypothesis which is consistent with the information of the error trial would be facilitated by the redundant information of the complementary stimulus being present along with the regular stimulus.

The design of the experiment assumed that the subjects would be stable with respect to their strategies by the end of the three practice problems. The performance measures were not constant, but this may have been due to changes in efficiency and error rates within the processing of a constant strategy. Toward the end of the sixteen problems, performance appeared to be approaching an asymptote. Solution probability was approaching one very closely -- again indicating that the task was not very complex or difficult. Also, Figure 6 shows a general decrease in the number of errors over problems.

There were two sources for the problem number effects on the general measures of performance. Practice accounted for the general rise of solution probability and decline of number of errors. This is further indicated by the increases in probability of solution for each dimension over occurrences shows in Figure 9 above. The different stimulus orders used on different problems account for almost all of the effects of problem number on trial of last error in Figure 8. If all of the information available from each stimulus sequence were processed by a subject, the four orders made the problems logically solveable as follows: order I, trial 4; order II, trial 3; order III, trial 3; order IV, trial 5. In addition, the varying difficulties of each dimension (from .65 to .90 probability of solution) and interproblem dependencies seem to have had an effect on the early problems.

The fixed number of trials methodology was a very successful technique. It supplied a large body of data in a short time; it was about the only practical way to obtain sufficient data from one subject to allow valid individual analyses. The length of the fixed trials problems was appropriate; most of the problems were solved in the interval alloted. The increase in mathematical complexity was not difficult to resolve for the fixed trials procedure.

The application of the sequential Bayesian evaluation technique to the present data was far from optimal. However, it was a large step above traditional analyses. There was no uncertainty or vagueness about the value of the Bower and Trabasso model in describing or predicting the observed behavior. Although no greater detail or accuracy was needed to form valid conclusions from this experiment, several areas of the technique are subject to improvement.

The four separate dimensional analyses per subject could be reduced to a single index of model fit to the individual. Four estimates of the model parameter were used. By ascertaining the dependencies between dimensions and their parameters, estimates could be made that are substitutable into a single evaluation expression for all four dimensions including all sixteen problems. The most obvious step that would need to be taken is to estimate the four parameters simultaneously in such a way that they sum to one. Given these estimates, it could perhaps be assumed that the parameter for a dimension is also applicable when the dimension is irrelevant. In this case the assumption of equally probable irrelevant dimensions involved in a non-final hypothesis selection following an error could be eliminated. Reducing the number of assumptions in a model is generally accepted as increasing the value of the model.

The sequential Bayesian evaluation technique used in this paper has great value. It is completely general, as long as there is more than one data point per subject. It can be applied to any well defined model in any area with no question of validity or comparability of its results with results from other models. It was efficient to apply to the data and faster to arrive at conclusions than other methods. No further information can be used from the data to evaluate a model; the technique supplies the probability of the model under evaluation at any point in the space on which the model is defined.

REFERENCES

REFERENCES

Atkinson, R., Bower, G., and Crothers, E. An Introduction to Mathematical Learning Theory. New York, John Wiley and Sons, Incorporated, 1965.

Blackwell, D. and Dubins, L. Merging of opinions with increasing information. Annals of Mathematical Statistics, 1962, 33, 882-886.

Edwards, W., Lindman, H., and Savage, L. Bayesian statistics for psychological research. Psychological Review, 1963, 70(3), 193-242.

Parzen, E. Modern Probability Theory and Its Applications. New York, John Wiley and Sons, Incorporated, 1960.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

MICHIGAN STATE UNIVERSITY
Computer Institute for Social Science Research
East Lansing, Michigan

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
A Sequential Bayesian Model Evaluation Technique Applied To A Model Of Concept Identification

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)
Gary H. Mendelsohn

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
30-December 1970	54	4

8a. CONTRACT OR GRANT NO. F44620-69-C-0114 (THEMIS)

b. PROJECT NO. 7921

c. 61102F

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)
70-5

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AFOSR-70-2883TR

10. DISTRIBUTION STATEMENT
Approved for public release;
distribution unlimited.

11. SUPPLEMENTARY NOTES
TECH, OTHER

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research
1400 Wilson Boulevard (NL)
Arlington, Virginia 22209

13. ABSTRACT

The determination of the empirical validity of mathematical models in the social sciences is a vexing problem. This report deals with a methodology for evaluating the goodness of fit of mathematical models which allows the researcher to utilize all of the information in each individual data protocol. This is accomplished by means of an iterative application of Bayes' Theorem, The technique is first explicated abstractly and is then applied to a particular mathematical model.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Bayes Theorem						
Concept Identification						
Empirical Validity						
Evaluation						
Goodness of Fit						
Mathematical Model						

UNCLASSIFIED

Security Classification